Classification du Cancer du Sein

Techniques d'Apprentissage Artificiel

Bademba SANGARE

Université Paris 8
Master 1 Informatique et Big Data

10 décembre 2025

Professeur : Rakia JAZIRI

Plan de la Présentation

Contexte et Définitions

Contexte : Apprentissage Supervisé

Nous disposons d'un jeu de données **étiqueté** (nous connaissons le diagnostic réel). Le but est d'entraîner un modèle à prédire cette étiquette pour de nouveaux patients.

Les Classes (Cibles)

- Maligne (M)
  - = **Malade**
  - = Positif (1)
- Bénigne (B)
  - = **Non Malade** (Sain)
  - = Négatif (0)

Objectif

- Le modèle doit pouvoir dire avec certitude si un patient est **Maligne** ou **Bénigne** en analysant les caractéristiques de la tumeur.

Le Jeu de Données (Dataset)

Source : Kaggle - Breast Cancer Wisconsin (569 patients)

Répartition des patients :

- 357 Bénignes (Non Malades)
  - Soit **62.7%** du total
- 212 Malignes (Malades)
  - Soit **37.3%** du total

Exemples de caractéristiques :

- `radius_mean` : Rayon moyen
- `texture_mean` : Texture
- `perimeter_mean` : Périmètre

ID	Diagnosis	Radius	Texture	...
842302	M	17.99	10.38	...
842517	M	20.57	17.77	...
84300903	M	19.69	21.25	...

Pipeline de Traitement

- **Split Train/Test (80/20)** :
  - **80% Entraînement** : Pour que le modèle apprenne.
  - **20% Test** : Données cachées pour évaluer la performance réelle.
- **Stratification** : On garde les mêmes proportions de malades dans les deux groupes.
- **Normalisation** : Indispensable pour mettre les données à la même échelle.

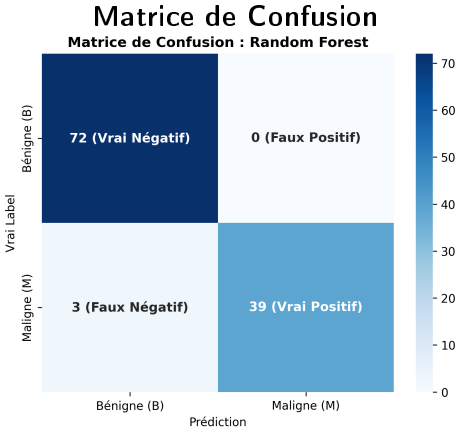
Les 3 Algorithmes Étudiés

- **Random Forest (Focus)** :
  Une "forêt" de 100 arbres de décision. C'est notre modèle principal car il est robuste et stable.

- **SVM (Support Vector Machine)** :
  Cherche la "frontière" optimale entre les deux groupes (Malades et Sains).

- **KNN (K-Nearest Neighbors)** :
  Regarde les **5 voisins** les plus proches. Simple mais sensible aux valeurs aberrantes (outliers).

Détail du calcul (Random Forest)



Données : Vrais Positifs=39, Vrais Négatifs=72, Faux Positifs=0, Faux Négatifs=3

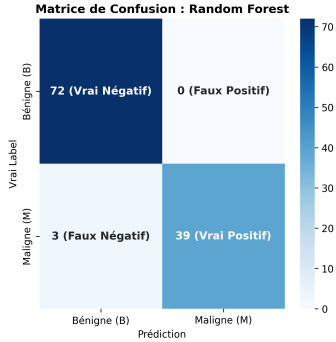
1. Accuracy
L'accuracy, c'est simplement le nombre de bonnes prédictions divisé par le total : 39 plus 72 sur 114 égale 97,37%.

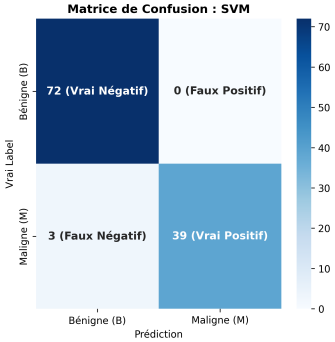
2. Precision
La precision, c'est les vrais positifs sur les vrais positifs plus les faux positifs : 39 sur 39 plus 0 égale 100%.

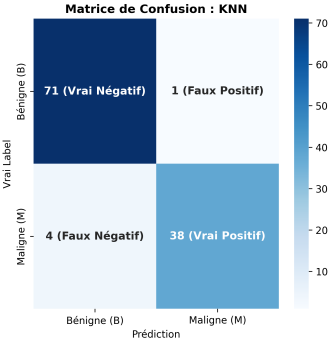
3. Recall
Et le recall, c'est les vrais positifs sur les vrais positifs plus les faux négatifs : 39 sur 42 égale 92,86%.

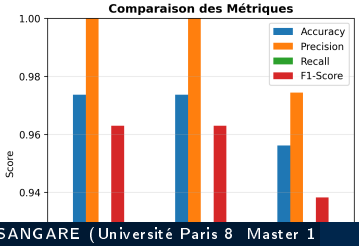
3 Faux Négatifs = point critique à améliorer
0 Faux Positifs = aucune fausse alerte

Comparaison Visuelle des Modèles









Performance Globale

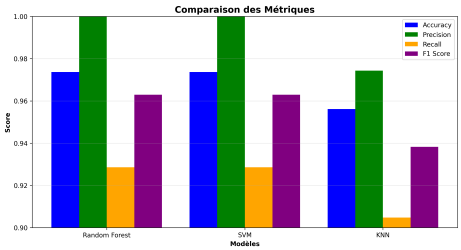


Tableau des Scores

Modèle	Acc.	Prec.	Rec.	F1	AUC
Random Forest	0.9737	1.0000	0.9286	0.9630	0.9929
SVM	0.9737	1.0000	0.9286	0.9630	0.9947
KNN	0.9561	0.9744	0.9048	0.9383	0.9823

Pourquoi ces résultats ?

- **RF et SVM (Top)** : Ils sont excellents pour gérer la complexité et les nombreuses dimensions (30 features) sans sur-apprendre.
- **KNN (En retrait)** : Basé sur la distance brute, il souffre plus du bruit ou des données isolées (outliers).

Conclusion

Bilan

- **Random Forest** est le modèle retenu.
- Il offre une précision de 100% (aucune fausse alerte).
- Le Recall de 92.8% signifie qu'on a raté 3 cas sur 42 : c'est le point d'amélioration prioritaire pour une application médicale réelle.

Merci de votre attention.